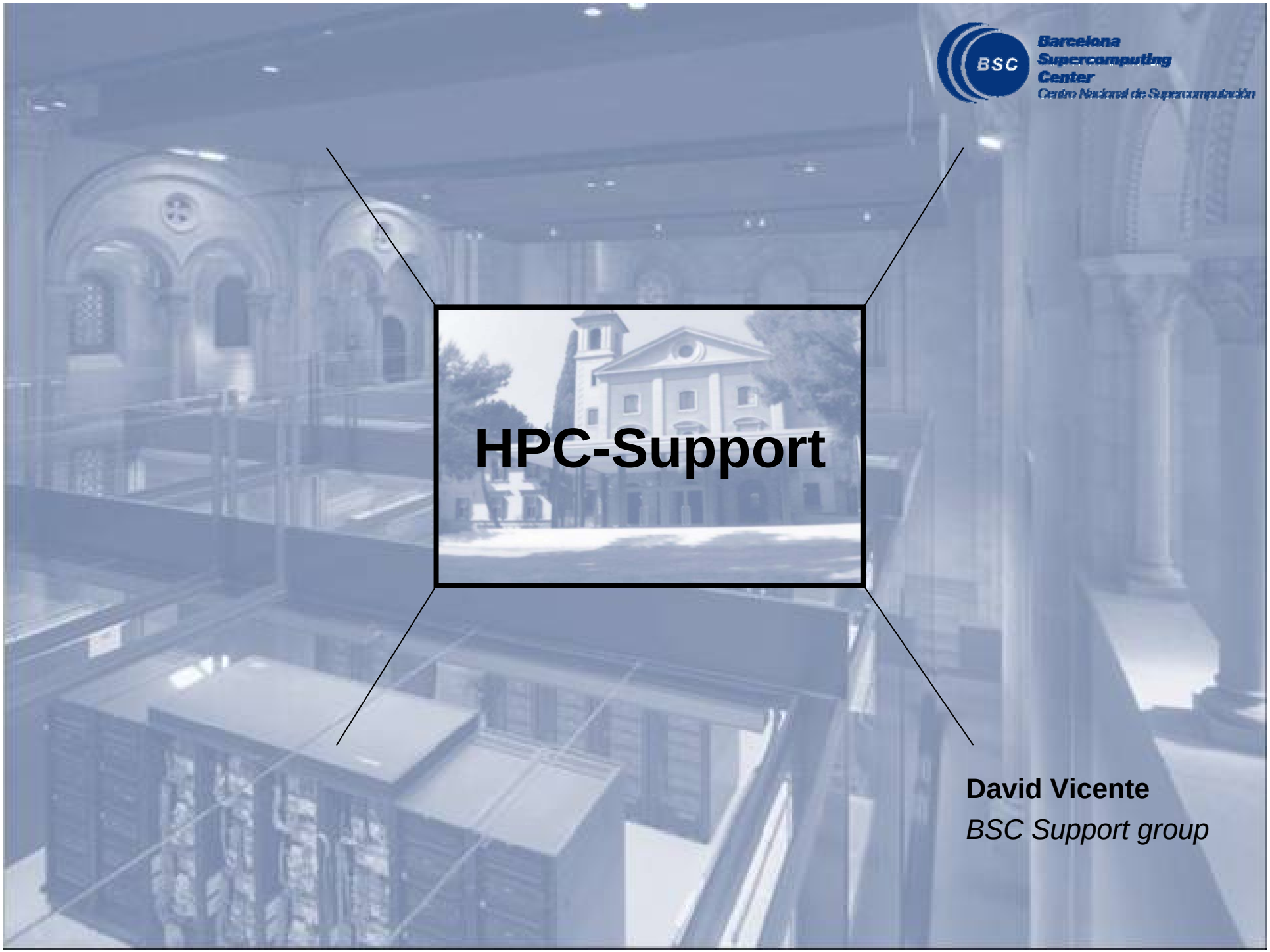


David Vicente
BSC Support group

Indice

- Buenas practicas para los usuarios
- Incidencias más comunes
- Información accesible por los usuarios

Puntos principales a nivel técnico de la arquitectura de las máquinas que componen la RES :

- Filesystems
- Sistemas de transferencia
- Sistema de colas
- Redes de conexión
- nodos
- Software

Filesystems

En el caso de MareNostrum disponemos de unos 370 TB de espacio de disco que está distribuido de la siguiente forma :

- /GPFS (Global Parallel Filesystem) : esta montado sobre la gigabit y es compartido por todos los nodos de la maquina.

/gpfs_scratch	**64T**
/gpfs_apps	**9.1T**
/gpfs_home	**19T (datos duplicados)**
/gpfs_projects	**73T**

- Disco local JS21 : disco SAS montado como /scratch en cada uno de los nodos de la máquina. En este mismo disco también se monta la partición de swap del blade(4GB).

/scratch 36G (solo 30 GB disponibles)

- Filesystems y uso (extraído del manual de usuario)
  - **/gpfs/home** :The home directory is to store the executables, own developed sources and personal data.
  - **/gpfs/projects:** This space is intended to store data that needs to be shared between the users of the same group (inputs, restarts …)
  - **/gpfs/scratch** : this filesystem is used only for temporal data during the execution of your job. Then it should be removed or transferred to a permanent filesystem.
  - **HSM** - Hierarchical Storage Management : This is the Hierarchical Storage Management (HSM) filesystem. HSM means that the filesystem moves files currently not used to tape. It is intended to archive LARGE files results and restarts

- Eficiencia de **GPFS** según su uso :
  - Al aumentar el numero de ficheros en un directorio, la eficiencia de las operaciones contra el filesystem, se reduce de forma contundente.
  - Programas con I/O masivo en bloques reducidos no funcionan de forma adecuada sobre /GPFS
  - Tamaño de los ficheros y block size

Filesystem	***Block size***
/gpfs/scratch	1MB
/gpfs/projects/	1MB
/gpfs/home/	256KB
/scratch (local)	4 KB

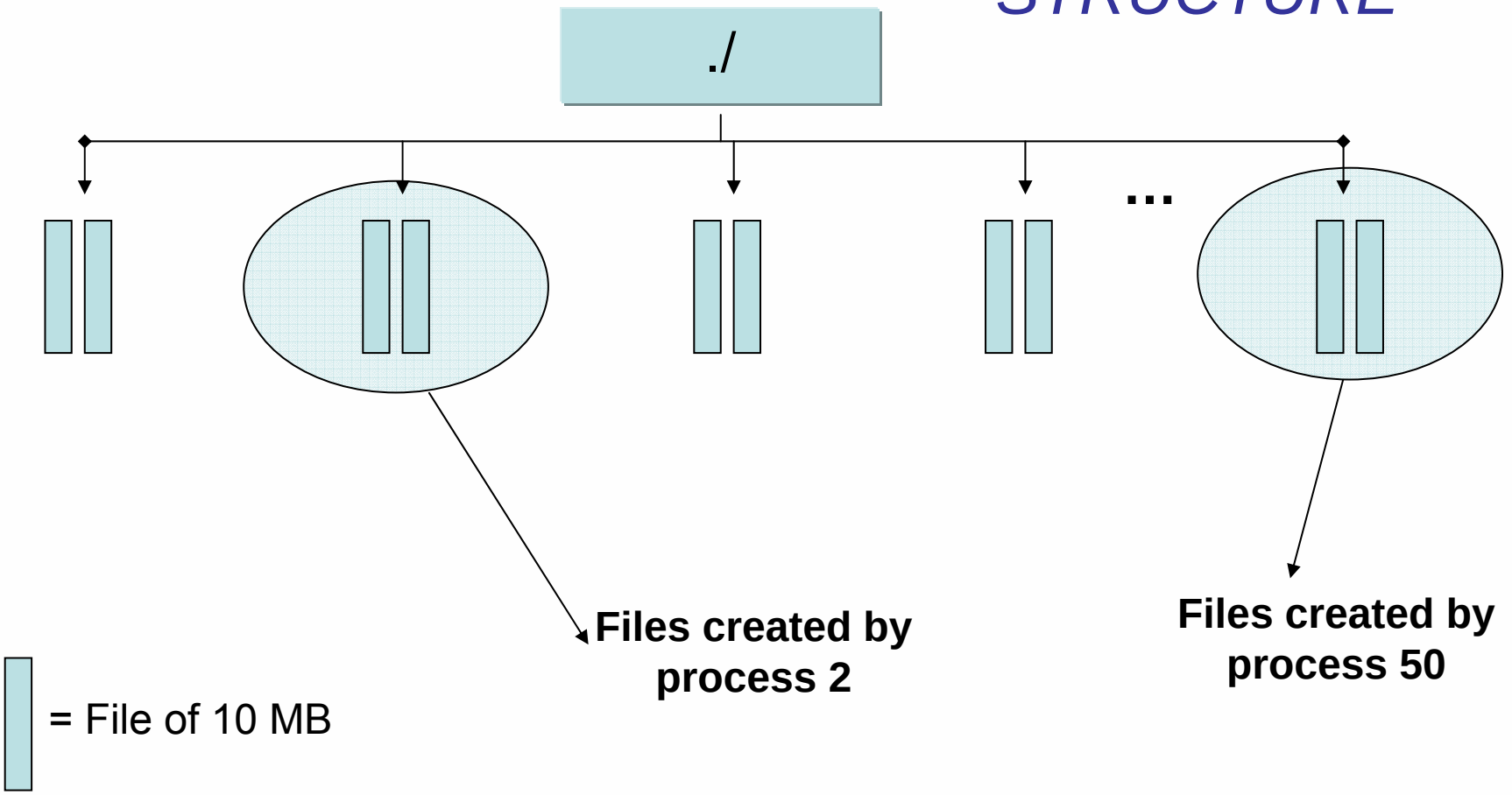
GPFS

Como mejorar el I/O de un proceso de usuario :

- Problemas con la creación de gran numero de ficheros en el mismo directorio :

2 files x 10 MB x 50 processes = 1 GB

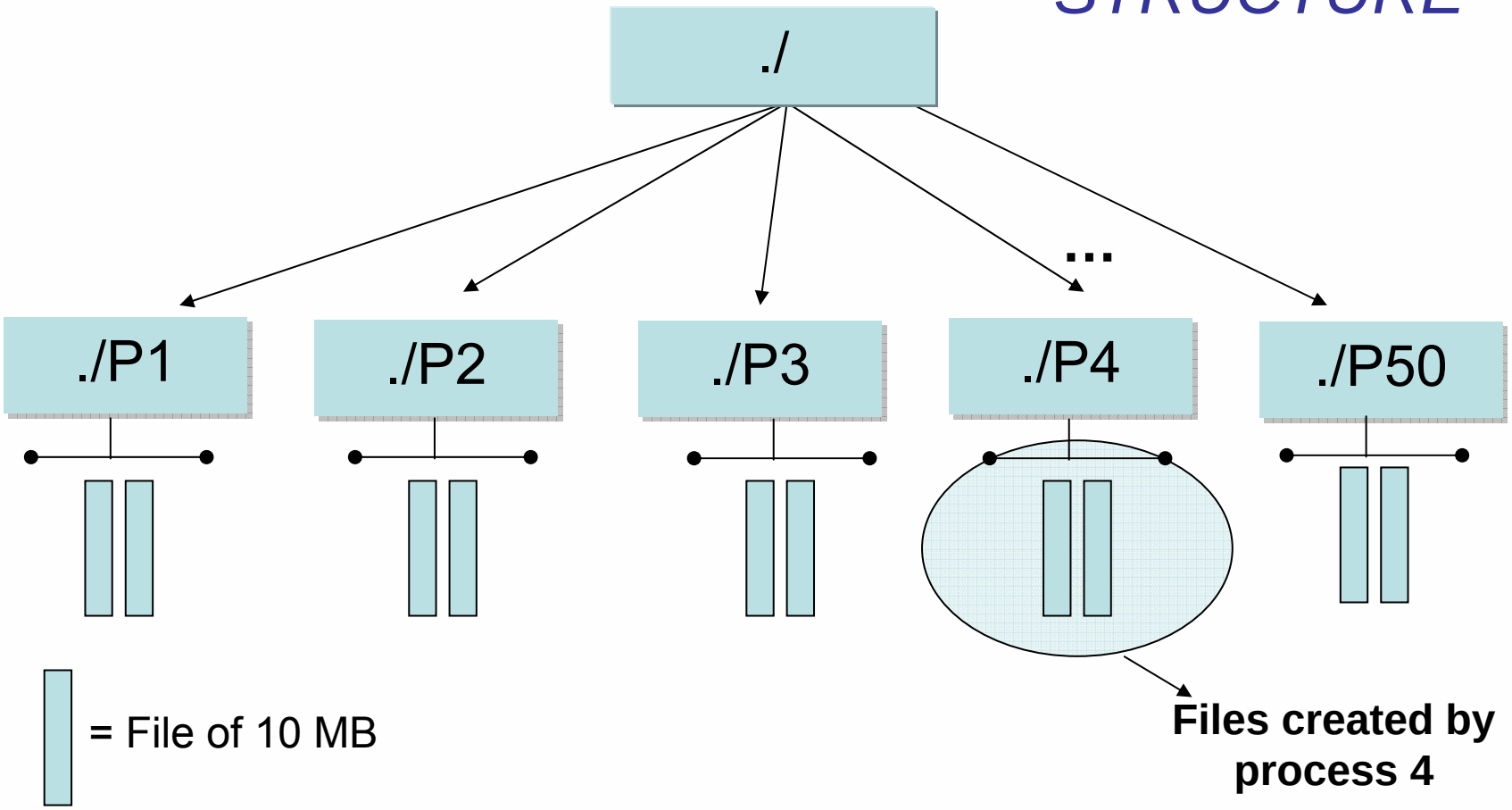
FLAT STRUCTURE



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

2 files x 10MB x 50 processes = 1 GB

TREE STRUCTURE



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

GPFS

- Files creation with a directory structure :

  50 CPUS:

  TREE MODE
  100 Files of 10485760 bytes written in 7.89267 seconds
  Agregated BW: 132.854 MBps

  FLAT_MODE
  100 Files of 10485760 bytes written in 45.3628 seconds
  Agregated BW: 23.1153 MBps

GPFS

Como mejorar el I/O de un proceso de usuario :

- Problemas con la creación de gran numero de ficheros en el mismo directorio :
- Problemas con operaciones masivas con bloques de tamaño reducido :
  - A ser posible, usar el disco local /scratch

GPFS

Como mejorar el I/O de un proceso de usuario :

- Disco Local
  - Al ser un disco dedicado para cada nodo, aumenta la eficiencia en trabajos con I/O masivo a nivel local. Por lo tanto es recomendable usarlo, siempre que los ficheros sean temporales y no se necesiten como resultado del trabajo.
  - Hay que tener en cuenta que actualmente no hay un proceso de limpiado del filesystem después de acabar los jobs. Esto dificulta su uso a nivel de usuario.

GPFS

Como mejorar el I/O de un proceso de usuario :

- Procesos dedicados a I/O
  - Para ciertos programas MPI, cuando la escalabilidad se ve muy afectada por el I/O, es posible dedicar ciertos procesos únicamente para las tareas de I/O, mejorando de forma substancial el rendimiento y escalabilidad.

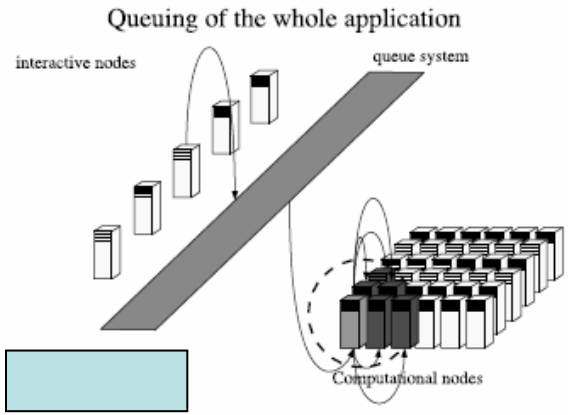
HSM

- Disponible en los login5, login6, login7 y login8.
- No esta montado en todos los nodos de la máquina. Los procesos ejecutados en el sistema de colas no pueden acceder a sus datos.
- Debido a que ciertas operaciones de I/O deben acceder a cintas para recuperar los datos y este proceso es costoso, se han creado ciertos comandos para facilitar el proceso :
  - hcp, hmv, htar
- Se recomienda tener ficheros de gran tamaño para mejorar la eficiencia al recuperar los datos desde cinta.

Sistema de transferencia

- Seguridad y transferencia
  - Debido a medidas de seguridad MN no permite conexiones salientes desde los logins de nodos, esto hace que todas las transferencias deban iniciarse siempre desde máquinas externas.
- Sistemas de transferencias disponibles en los sites de la RES :
  - SSH
    - Tanto los datos como los passwords van encriptados
    - El proceso de encriptación consume CPU y limita de forma considerable el ancho de banda.
  - FTPS
    - Solo el password se encripta, los datos van en claro.
    - Al no encriptar los datos, el consumo de recursos es mínimo, esto también aumenta el ancho de banda total en un 20% respecto SSH
    - Las conexiones FTPS deben hacerse contra la máquina mn1-ftps.bsc.es (ver manual de usuarios sección 4.2)

- Actualmente tenemos dos sistemas de colas disponibles en los sites de la RES :
  - **LoadLeveler** : Madrid, Zaragoza
  - **Slurm+MOAB**: Malaga, LaPalma, BSC, Cantabria, Valencia



LoadLeveler

- Sistema de colas proporcionado por IBM
- http://www-03.ibm.com/systems/clusters/software/loadleveler.html
- Problemas :
  - No gestiona el paralelismo (MPI) de los trabajos paralelos, esto provoca problemas en la gestión de recursos con ejecuciones de más de 512 cpus.
  - Su sistema de prioridades y fair-sharing es más limitado que el proporcionado por MOAB

SLURM / MOAB

- **SLURM** desarrollado por Lawrence Livermore National Laboratory (http://www.llnl.gov/linux/slurm/)
- Es Open-Source y esta especialmente diseñado para clusters Linux de cualquier tamaño.
- SLURM es capaz de controlar el paralelismo a nivel de aplicación (srun)
- **MOAB** aplicación que gestiona el cluster a nivel de prioridades, monitorización, reservas, fair-sharing, etc . Esta desarrollada por la empresa “Cluster Resources” .

Limitaciones a nivel de usuario para el lanzamiento de trabajos a colas :

- Mínimo tiempo por trabajo (10 min excepto jobs de debug).
- Máximo numero de trabajos (100 por usuario)
- Numero máximo de cpus por trabajo
  - 1024 (loadleveler)
  - 10240 (slurm) *
- Máximo tiempo por trabajo :
  - Horas con prioridad → 72 horas
  - Horas sin prioridad → 24 horas

Redes a nivel de usuario :

- Myrinet :
  - Proporcionada por Myricom
  - Ancho de banda -> 2Gb/s punto a punto
    (cada nodo dispone de una conexión myrinet)
  - Latencia minima (3,5 µs)
  - Solo para aplicaciones MPI
- Gigabit :
  - Ancho de banda 1Gb/s por Blade Center (14 blades)
  - Solo para conexiones SSH y GPFS

BLADE JS21

- Procesador PPC970-MP
  - 4 procesadores por blade (2 dual core chips)
  - 2.3 Ghz
  - 4 Flop por ciclo (9.2 GFlops)
  - Soporta instrucciones ALTIVEC (VMX)
    - Solo en precisión simple

- Memoria
  - 8 GB por nodo
  - 800 MB usados por sistema
  - 7.2 GB para aplicaciones de usuario (1.8 GB por proceso)

BLADE JS20

- Procesador PPC970-FX
  - 2 procesadores por blade
  - 2.2 Ghz
  - 4 Flop por ciclo (4.8 GFlops)
  - soporta instrucciones ALTIVEC(VMX)
    - Solo precisión simple
- Memoria
  - 4 GB por nodo
  - 800 MB usados por S.O y GPFS
  - 3.2 GB disponibles para aplicaciones de usuario (1.6 GB por proceso)

- Software:
  - compiladores
    - GCC/G77 (v3.3.3)
    - GFORTRAN (v 4.0.1, v 4.1.2) /gpfs/apps
    - XLC/XLF (IBM compilers, recomendados)
  - MPICH-GM
    - Implementación de MPI sobre Myrinet mediante el protocolo GM
    - mpicc/ mpif77/ mpif90
  - MPICH-MX *
    - Implementación de MPI sobre Myrinet mediante el nuevo protocolo MX, por ahora no está en producción en ningún site de la RES.

- Algunos puntos importantes para hacer porting de aplicaciones:
  - Compilar el código con los compiladores de IBM
    - Flags (-qtune=ppc970 –qarch=ppc970 –qcache=auto –qstrict –O3)
    - OMP support (-qsmp=omp)
  - Utilizar las librerías más optimizadas para nuestra arquitectura :
    - FFTW con ALTIVEC
    - GOTO-BLAS
    - ESSL
    - MASS
    - Myrinet MPI library

Performance study

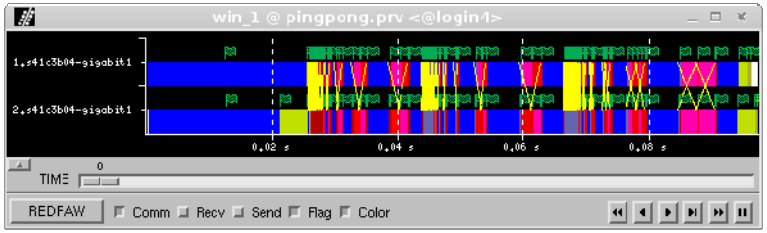
- IHPCT (IBM High Performance Computing Toolkit)
- Interposing PAPI counters to MPI calls
- MPIP
- PERFMINER
- PERDMF (TAU)
- Paraver (BSC-CNS)

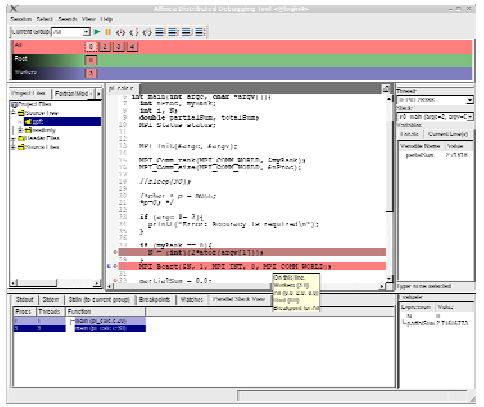
Profiling

- GNU profiler (gprof,kprof)
- Xprofiler
- Paraver (BSC-CNS)

Debugging

- DDT (parallel debugger)
- GDB





Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Máquina de memoria compartida

- Actualmente, el BSC dispone de una máquina de memoria compartida que puede ser útil para que algunos usuarios puedan hacer el pre-proceso o post-proceso de los datos generados en MN.

  - Características :
    - **SGI Altix 4700**
    - **Peak Performance of 819,2 GFlops**
    - **64 Itanium2 dual core processor at 1.6 GHz**
    - **512 GB of main memory**

Incidencias

Descripción de las Incidencias o errores típicos que pueden aparecerle a un usuario de la RES, en este caso hemos ordenado los problemas según estas 4 categorías :

- Compiladores
- MPICH-GM
- Sistema de colas
- Otros

ERROR

> mpif77 -o speed.fftw.mpi speed.fftw.mpi.o mpiwrap.o -L/gpfs/apps/FFTW/32/lib -lrfftw -lfftw -lrfftw_mpi -lfftw_mpi

ld: skipping incompatible /gpfs/apps/FFTW/32/lib/librfftw.a when searching
for -lrfftw
ld: cannot find -lrfftw

CAUSA :

Se están intentando linkar objetos con 32 y 64 bits

SOLUCION :

Recompilar todos los objetos con el mismo formato (-q32, -q64)

ERROR

> mpif77 -o speed.fftw.mpi speed.fftw.mpi.o mpiwrap.o -L/gpfs/apps/FFTW/32/lib -lrfftw -lfftw -lrfftw_mpi -lfftw_mpi

/scratch/tmp/ccXTvT4j.o(.text+0x6c): In function `MAIN__':

undefined reference to `rfftw3d_f77_create_plan__'

CAUSA :

Las llamadas a Fortran compiladas con los compiladores de IBM no tienen el underscore final, típico de G77

SOLUCION :

Recompilar con la opción –qextname o redefinir estas llamadas de la forma adecuada –WF,-Dfortran_call=fortran_call_

Consulta :

¿Cómo puedo saber si mi binario paralelo usa la myrinet para comunicaciones?

SOLUCION :

```
> ./test_cpu
<MPICH-GM> Error: Need to obtain the job magic number in GMPI_MAGIC !


> ldd test_cpu
        time64.so => /home/bsc99/bsc99349/TEST-GM/time64.so (0x0000008000030000)
        libdl.so.2 => /lib64/libdl.so.2 (0x0000008000050000)
        libmpich.so.1.0 => /opt/osshpc/mpich-gm/64/lib/shared/libmpich.so.1.0 (0x0000008000064000)
        libgm.so.0 => /usr/lib64/libgm.so.0 (0x000000800026d000)
        libpthread.so.0 => /lib64/tls/libpthread.so.0 (0x00000080002a7000)
        libm.so.6 => /lib64/tls/libm.so.6 (0x00000080002d3000)
        libc.so.6 => /lib64/tls/libc.so.6 (0x000000800035b000)
        /lib64/ld64.so.1 (0x0000008000000000)

>  nm -A test_cpu | grep -i mpi
```

Consulta :

¿Cómo puedo linkar con las librerías PESSL?

SOLUCION :

```
LIBS_IBM="-L/opt/ibmcmp/xlf/10.1/lib64 -R/opt/ibmcmp/xlf/10.1/lib64 -L/opt/ibmcmp/xlsmp/1.6/lib64 \
-R/opt/ibmcmp/xlsmp/1.6/lib64 -lpesslgm -lblacsgm -lm -lessl -lxlf90_r -lxlomp_ser -lxlfmath -lmpichfarg"

ALL :  example

example : example1.o
	$(MPIF77) example1.o $(LIBS) -o example

example1.o : example1.f
	$(MPIF77) $(FFLAGFS) example1.f -c

clean :
	rm -f *.o example
```

ERROR

Use of uninitialized value in subroutine entry at
/opt/osshpc/mpich-gm/64/bin/mpirun.ch_gm.pl line 862.
Bad arg length for Socket::inet_ntoa, length is 0, should be 4 at
/opt/osshpc/mpich-gm/64/bin/mpirun.ch_gm.pl line 862.

CAUSA :

Se llama MPIRUN sin machinefile o esta está vacía.

SOLUCION :

Usar la llamada estándar en el script de colas :

mpirun –np $NPROCS –machinefile $LL_MACHINE_LIST ./binario

ERROR

Se produce un segmentation fault en la creación de una automatic array de Fortran.

CAUSA :

El limite de stack es inferior al necesario para el ejecutable, en este caso esta limitado a 1GB en los nodos de calculo.

SOLUCION :

Aumentar el limite de stack o recompilar la aplicación con -qsave

ERROR

[13] MPI Abort by user Aborting program !

[13] Aborting program!

CAUSA :

La MPI task con ID=13 ha abortado con un MPI_Abort

ERROR

```
Permission denied (publickey,keyboard-interactive).
Permission denied (publickey,keyboard-interactive).
Permission denied (publickey,keyboard-interactive).
Permission denied (publickey,keyboard-interactive).
Permission denied (publickey,keyboard-interactive).
Permission denied (publickey,keyboard-interactive).
Permission denied (publickey,keyboard-interactive).
Permission denied (publickey,keyboard-interactive).
Permission denied (publickey,keyboard-interactive).
```

CAUSA :

El SSH entre nodos ha fallado, esto solo pasa ejecutando con MPIRUN + LoadLeveler, en SLURM todo este proceso es llevado a cabo por srun.

Hay diversas causas que pueden provocar este problema algunas de ellas son :

- Permisos del directorio .ssh o HOME
- Claves SSH para acceder sin password entre nodos
- Problemas con GPFS en algún nodo involucrado en al ejecución

ERROR

El job acaba con un mensaje de error parecido al siguiente :

LoadLeveler :

Killed by signal 15.

Slurm :

srun: Force Terminated job

CAUSA :

El trabajo ha superado el wall_clock_limit

SOLUCIÓN :

Aumentar el wall_clock_limit del trabajo, modificando el parámetro dentro del script de colas

ERROR

s32c4b09-gigabit1
s32c4b09-gigabit1
ppc64
/usr/local/man:/usr/share/man:/usr/X11R6/man:/opt/gnome/share/man
news
no

CAUSA :

Es un bug con tcsh/csh y loadleveler

SOLUCIÓN :

Modificar el script para que use /bin/bash

Consulta :

¿Como puedo lanzar un trabajo MPI+OpenMP en la máquina ? Ejemplo 10 procesos MPI x 2 OMP (20 cpus en total)

SOLUCION SLURM:

```
#@total_tasks =10
#@ cpus_per_task =2
```

SOLUCION LoadLeveler:

```
#@total_tasks =10
#@blocking = 2
```

ERROR

> Signal 11 (segmentation fault)

CAUSA :

El proceso hace un acceso a memoria invalido

SOLUCION :

Ayudar al usuario a debugar su aplicación

- **Como hacer debugging de una aplicación en las máquinas de la RES**
  - Recompilar el código con la opción –g para poder proporcionar información sobre la línea exacta del programa donde se esta produciendo el problema.
  - Como los usuarios no tiene acceso directo a los nodos de computación, hay que proporcionarles algún script para que puedan debugar sus aplicaciones de forma remota.

- **Script para generar un core:**

Útil si sabemos con certeza la task_id que esta provocando el problema, esto es posible si después de varias ejecuciones el programa siempre falla en el mismo punto y el signal siempre proviene del mismo proceso :

```
> cat core.sh
#!/bin/bash
EXE=$@
if [ "$GMPI_ID" = 8 ]; then
        ulimit -c 1000000
        $EXE
else
$EXE
```

- **Script para debugar en colas con strace :**

  Evita el uso de GDB, sirve para saber exactamente que tipo de llamadas a sistemas están llevándose a cabo en cada punto del proceso.

```
> cat strace.sh
#!/bin/bash

EXE=$@

if [ "$GMPI_ID" = 0 ]; then
    strace -o file$$ $EXE
    #strace64 -o file$$ $EXE
else
    $EXE
fi
```

- **Script para debugar en colas con gdb :**

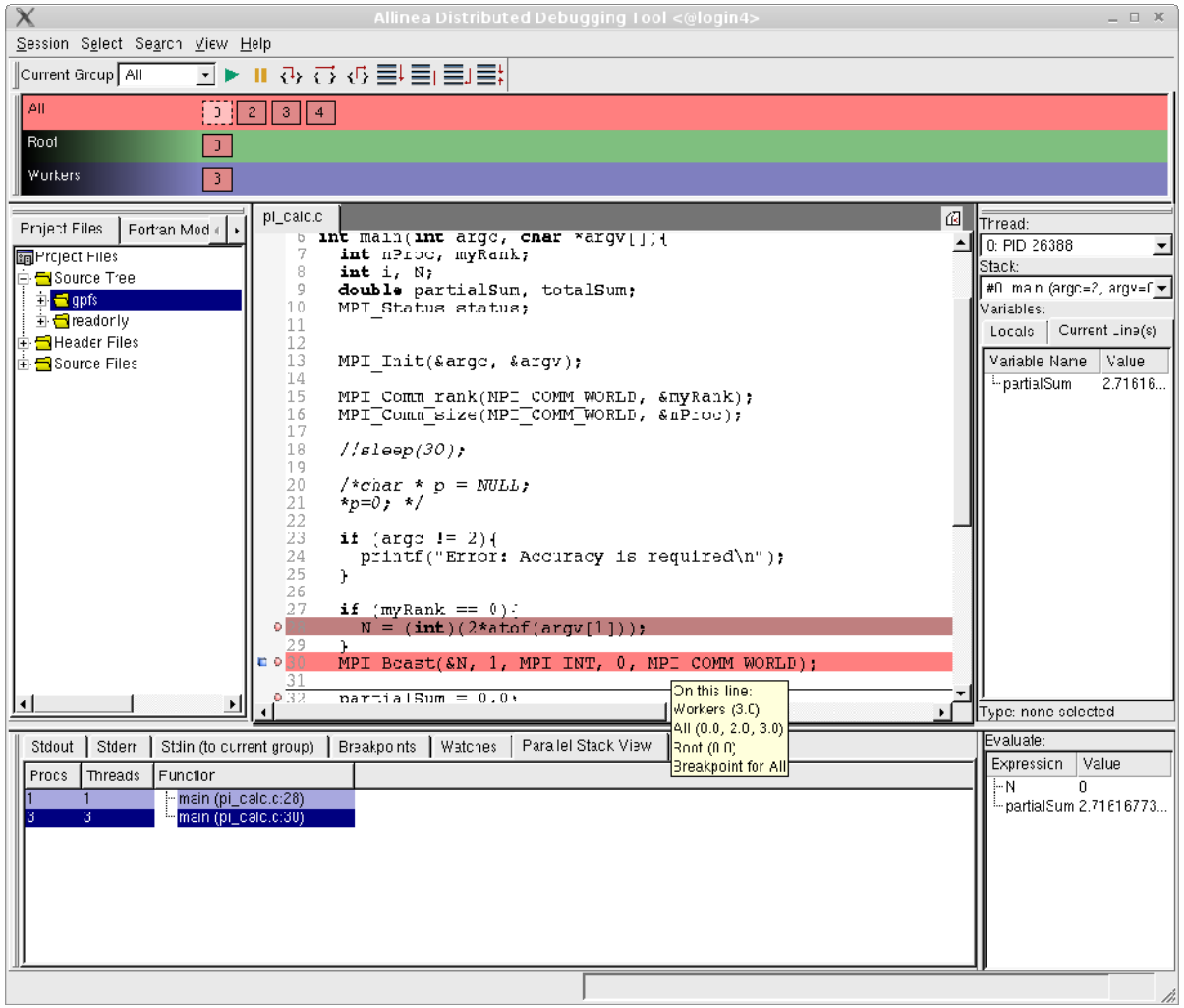
Sirve para ejecutar gdb en modo batch, de esta forma puede enviarse el job a colas sin necesidad de una sesión interactiva.

```
> cat strace.sh
#!/bin/bash

EXE=$@

if [ "$GMPI_ID" = 0 ]; then
    gdb –batch –args $EXE
    #gdb64 –batch --args $EXE
else
    $EXE
fi
```

- Debugging MPI con DDT:



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Información Usuarios

- **Disponible en la pagina web :**
  - **Listado de los proyectos que se están ejecutando actualmente en los diferentes sites de la RES:**

    http://www.bsc.es/plantillaA.php?cat_id=375

Listado proyectos

- RES
- Mobility programs
- Services
- System Architecture
- Available Software
- Current Activities
  - Astronomy, Space and Earth Sciences
  - Biomedicine and Life Sciences
  - Physics and Engineering
  - Chemistry and Material Science

Astronomy, Space and Earth Sciences

Title: Assimilation of High Resolution Precipitation Estimates in an Ensemble of Numerical Models
Leader: FRANCISCO J. TAPIADOR
Institution: UNIVERSIDAD DE CASTILLA-LA MANCHA en Toledo
Abstract:
The aim of this project is to evaluate the performance of a standard weather models used for numerical weather prediction (NWP) when three-dimensional high-resolution satellite-derived rainfall estimates are assimilated. Our working hypothesis is that a continuous assimilation of the satellite rainfall estimates will improve the forecasts of extreme events. This activity is supported by the CICYT project ASPRES: Asimilación 4-D de estimaciones de precipitación en un modelo numérico para la mejorar de la predicción de extremos hidrometeorológicos en España.

Title: Heisenberg Spin Glasses: Large Lattices at Low Temperatures
Leader: Victor Martin Mayor
Institution: Universidad Complutense de Madrid
Abstract:
Spin-glasses are disordered magnetic alloys with very unusual behaviour. Interest in spin glasses goes far beyond just this class of materials, however, because they provide a convenient model system from which to derive a unified theory of the glassy state (glass is one of the most common, yet mysterious states of condensed matter). Interest in spin glasses also extends outside of condensed matter physics because there has been a fruitful exchange of ideas between the spin glass community and researchers in other branches of science such as computer scientists working on combinatorial optimization. Simple analytical approximations predict that spin glasses have a finite temperature transition, but whether this actually occurs (or is an artifact of the approximations used) has remained unclear, at least for a variety known as "Heisenberg spin glasses". The advent of a new generation of simulation algorithms, finite-size scaling analysis and parallel computers have meant that, at last, we can make a serious attempt to fully understand the phase transition in Heisenberg spin glasses.

Title: The Marenostrum Numerical Cosmology Project: Grand Challenge simulations of structure formation in the Universe
Leader: Gustavo Yepes Alonso
Institution: Universidad Autonoma de Madrid
Abstract:
The MareNostrum Numerical Cosmology Project (http://astro.ft.uam.es/marenostrum) is an international collaboration with the aim of using the exceptional capabilities of the MareNostrum supercomputer to carry out grand challenge cosmological simulations of the formation of galaxies, clusters of galaxies and the large scale structures in the universe. The scientific objective is to understand the physical processes that were involved in the formation and evolution of these objects from initial conditions that are compatible with the early epochs of the Universe derived from the observations of the Cosmic Microwave Background radiation.
The MareNostrum Galaxy Formation Simulation is one of the largest SPH simulation of galaxy formation performed to date which uses more than 2 billion particles to represent the different components of the matter in the universe: dark matter, gas and stars. The evolution of the self-gravitating fluids is done by means of the N-body+SPH parallel MPI code GADGET-2. This simulation is also done with a different computational code, based

Información Usuarios

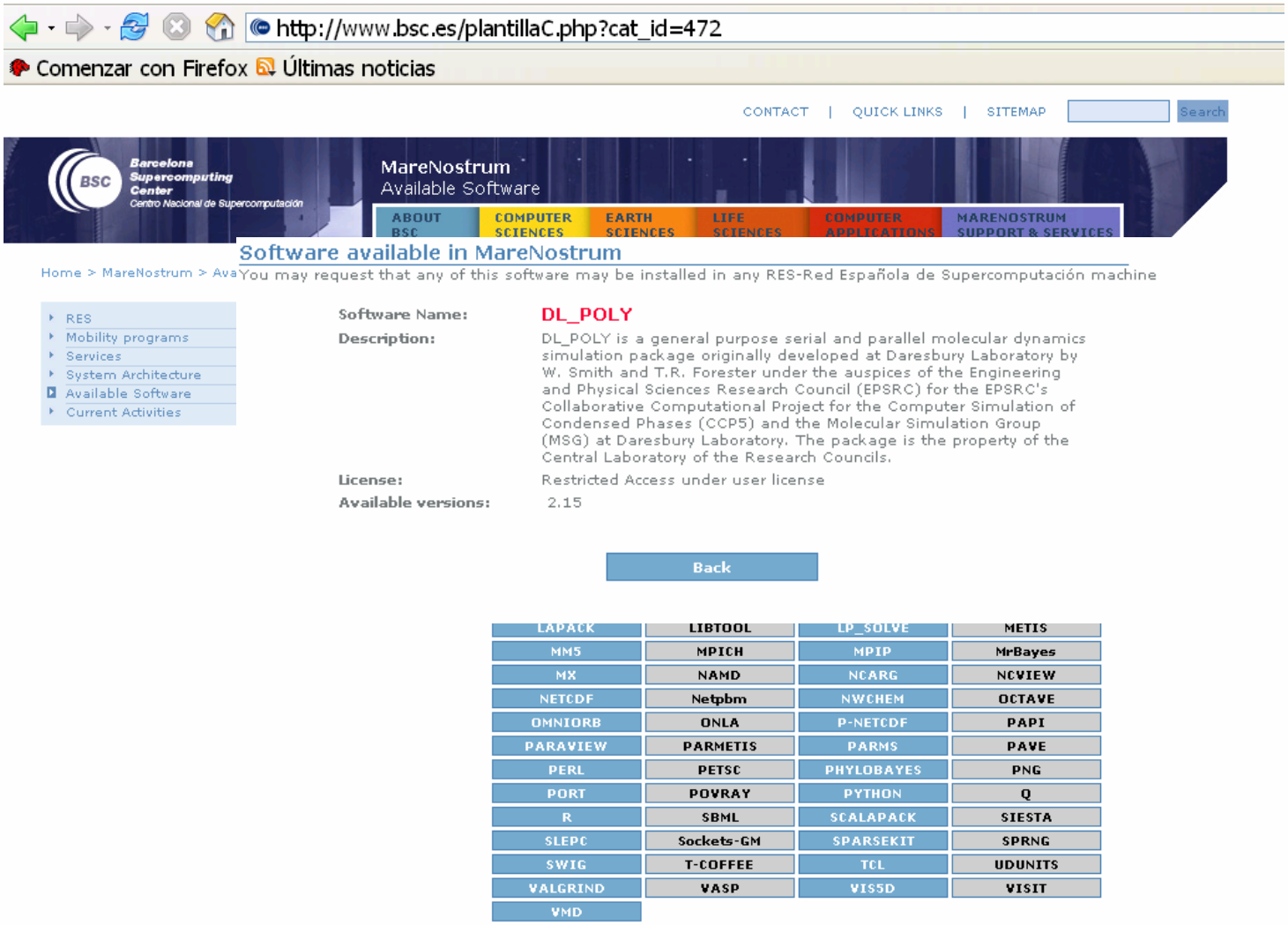
- **Disponible en la pagina web :**
  - Listado de los proyectos que se están ejecutando actualmente en los diferentes sites de la RES:

    http://www.bsc.es/plantillaA.php?cat_id=375

  - Listado de todas las aplicaciones disponibles en la RES:

    http://www.bsc.es/plantillaC.php?cat_id=472

Listado Aplicaciones



Información Usuarios

- **Disponible en la pagina web :**
  - Listado de los proyectos que se están ejecutando actualmente en los diferentes sites de la RES:

    http://www.bsc.es/plantillaA.php?cat_id=375

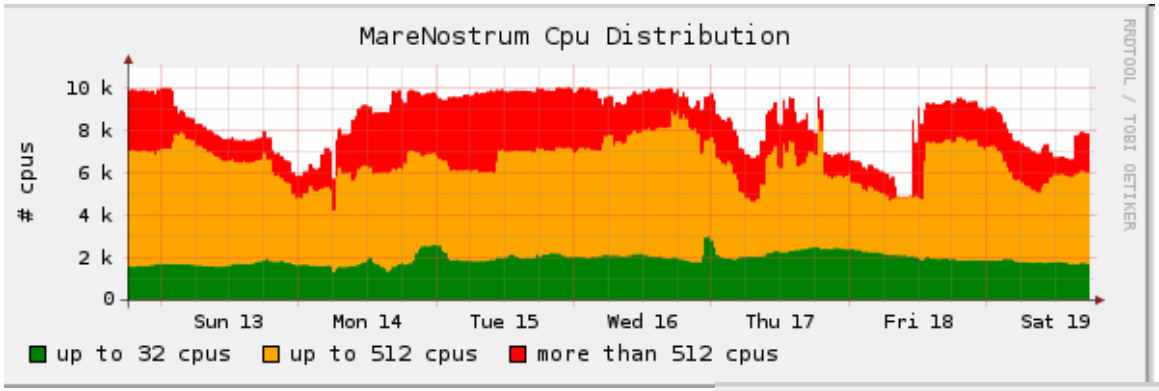
  - Listado de todas las aplicaciones disponibles en la RES:

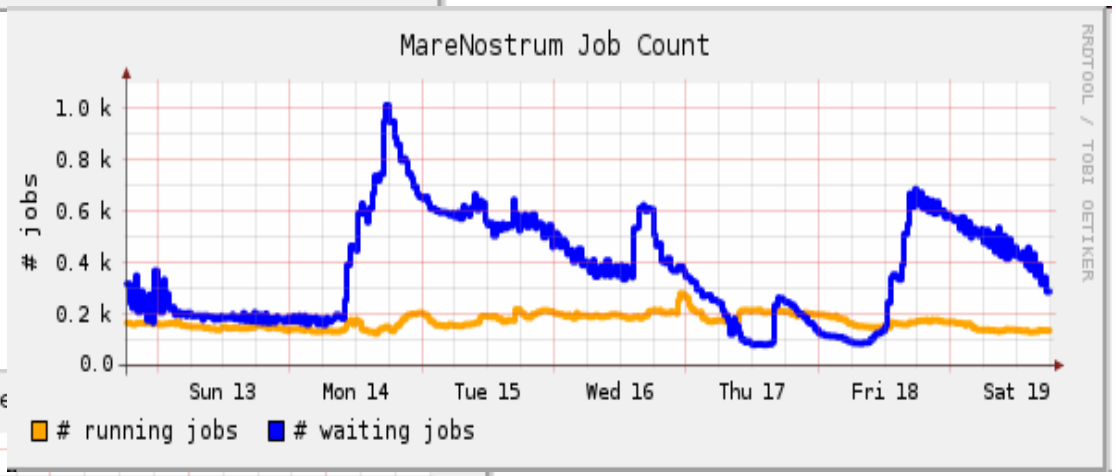
    http://www.bsc.es/plantillaC.php?cat_id=472

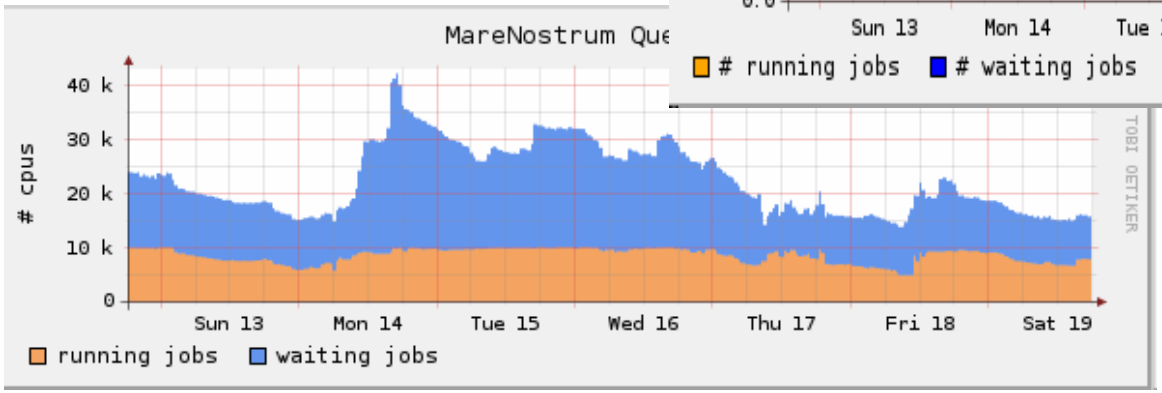
  - Gráficas de uso del sistema:

    http://www.bsc.es/plantillaA.php?cat_id=5

Información Usuarios





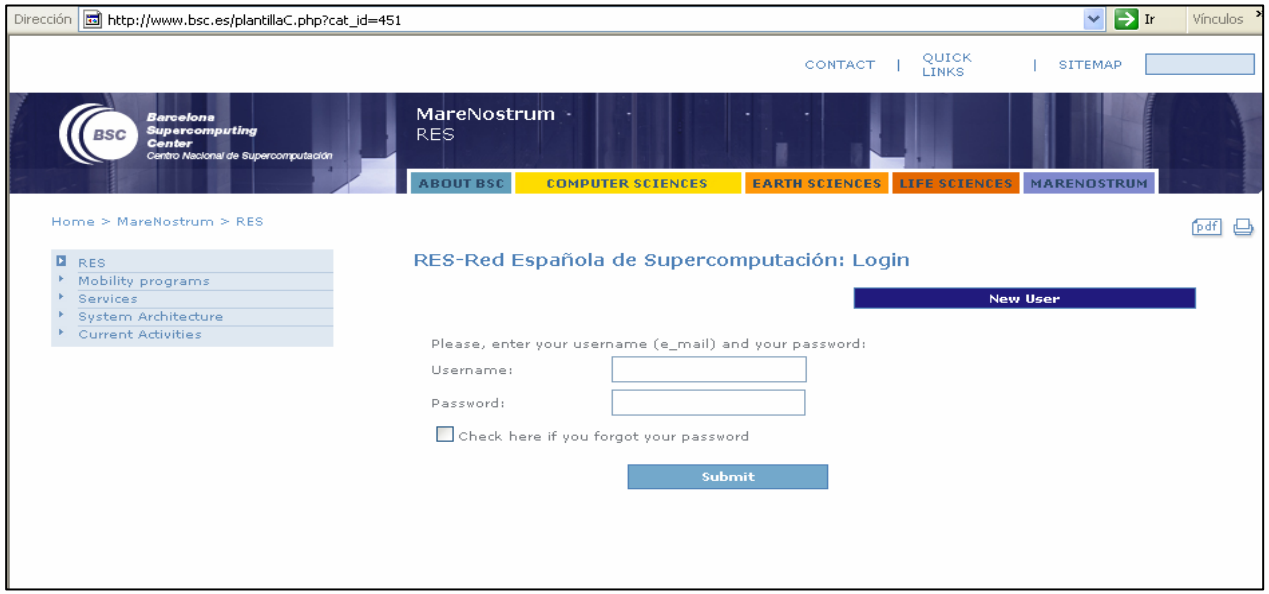


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Información de usuario RES-AREA

RES-AREA - Página de acceso :

http://www.bsc.es/RES



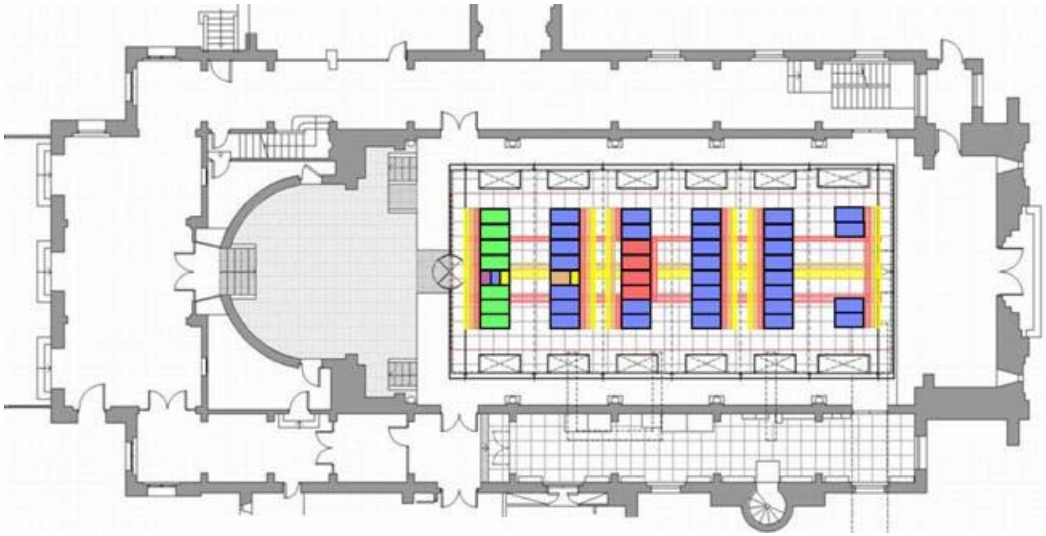
RES-AREA (servicios para los usuarios)

- Creación de nuevas cuentas
- Generación del formulario para peticiones de actividades de supercomputación
  - Nuevos proyectos
  - Continuaciones de proyectos anteriores.
- Consumo
- Redacción de informes técnicos sobre el uso de la máquina.
- Espació para información de marketing
- Acceso a la encuesta de satisfacción de usuarios

- Acceso (usuario estándar):
  - Username : e-mail
  - Password, enviado por e-mail al marcar la casilla:
    - Check here if you forgot your password

RES-AREA (vista de usuario)

Gracias !!



Barcelona Supercomputing Center
Centro Nacional de Supercomputación